Contributions of the Genome Sequence of *Erwinia amylovora* to the Fire Blight Community

A.M. Bocsanczy and S.V. Beer
Department of Plant Pathology
Cornell University
Ithaca NY 14853
USA

N.T. Perna, B. Biehl and J.D. Glasner
Animal Health and Biomedical Sciences
University of Wisconsin-Madison
Madison, WI 53706-1581
USA

S.W. Cartinhour, D.J. Schneider and G.A. DeClerck
USDA Plant, Soil and Nutrition Research
Cornell University
Ithaca NY 14853
USA

M. Sebaihia, J. Parkhill and S. Bentley
Pathogen Informatics Team
The Wellcome Trust Sanger Institute
Cambridge
UK

Keywords: microbial genomes, enterobacteria, Ea273, annotation, ASAP, plant pathogen

Abstract

The entire genome of *E. amylovora*, strain Ea273, was sequenced at the Sanger Institute. A manuscript reporting the total sequence and its annotation is in preparation by authors in three locations. The genome of Ea273 is ca. 3.8 mbp, which is small compared with other enterobacteria including non-pathogenic *E. coli* K-12, and it shows signs of erosion due to many pseudogenes and "lost" genes. It also has a limited number of homologs with proteins associated with the type three secretion system (T3SS) of other sequenced Gram-negative plant pathogens. Analysis of a preliminary annotation identified several potential novel virulence factors, including two novel T3SS islands. The genome sequence can be accessed at http://www.sanger.ac.uk/Projects/E_amylovora/. The preliminary annotation and sequence can be accessed on request and registration at: https://asap.ahabs.wisc.edu/asap/logon.php. The ASAP database, maintained at the University of Wisconsin, Madison, is an extensive compendium of information relative to microbial genomes, along with tools that facilitate comparisons of the genome of Ea273 with that of other sequenced bacteria. To increase the value of the genome to the fire blight community and others, additional biological annotation is solicited. Curation of the genome will be done both at UW-Madison by Nicole Perna and at Cornell by Ana Maria Bocsanczy. In the following few years, we hope that the genome sequence will facilitate greater understanding of fire blight at both the molecular genetic and practical disease control levels.

INTRODUCTION

Fire blight is a devastating disease of rosaceous plants. It affects all four subfamilies of the Rosaceae family, infecting about 200 species (Momol and Aldwinckle, 2000). However, the main economic importance of this disease is its effect on the apples and pears throughout the world. The Gram-negative bacterium *Erwinia amylovora* causes fire blight. There is no known plant host for *E. amylovora* outside the family Rosaceae. *E. amylovora* is the type species of the genus *Erwinia*, which was created to contain members of the *Enterobacteriaceae* that are associated with plants (Paulin, 2000).

E. amylovora is phylogenetically related to important animal and plant pathogenic enterobacteria, such as *Escherichia coli*, *Salmonella* spp., *Shigella* spp., *Yersinia* spp., *Pectobacteium carotovorum* and *E. chrysanthemi*. *E. amylovora* shares many biochemical and morphological characteristics with members of this family, suggesting that many survival and environmental responses could be conserved. However as a plant pathogen, *E. amylovora* has several features, like a T3SS pathogenicity island (Oh et al., 2005), that resemble those of phylogenetically more distantly related Gram-negative bacteria, such as

Pseudomonas syringae.

The complete genome of *E. amylovora* was recently sequenced at The Sanger Institute (www.sanger.ac.uk/Projects/Microbes). Strain Ea273 was selected for the sequencing, in part, because it is highly virulent on important hosts such as apple and pear. The sequence and the annotations were uploaded to the ASAP (A Systematic Annotation Package for community analysis of genomes). ASAP is a comprehensive database developed to store, update and distribute genome sequence data and annotation for all enterobacteria. The site (https://asap.ahabs.wisc.edu/annotation/php/ASAP1.htm) is a web interface that facilitates input and sharing of information by researchers distributed throughout the world. Currently, the ASAP database includes a comprehensive list of enterobacterial genome sequences, with their annotation, experimental data, when available and tools for comparison with public databases. ASAP supports three levels of users: public viewers, annotators and curators (Glasner et al., 2003). The genome sequence and the preliminary annotation are available upon request at the ASAP database. Access is limited to registered users, who can access it as annotators, and have agreed to abide with the ASAP data release policy. Upon completion of initial analyses, the Ea273 genome will be publicly available through ASAP for continuous updates of the annotation to reflect the growing knowledge of the community.

A manuscript reporting the sequence and a number of putative novel virulence factors and regions potentially associated with virulence and pathogenicity is in preparation at three locations. Among the more interesting results, we report the identification of two novel T3SS islands, related to animal pathogenic and endosymbiotic bacteria. The genome comparison with other closely related genomes in the Enterobacteriaceae suggests that *E. amylovora* has a compact genome. It is closely related to that of *E. coli* and *Salmonella* spp., in structure, but *E. amylovora* lacks coding regions found in many other enterobacteria that are clearly not required for its plant pathogenic lifestyle.

The sequencing of the genome of *E. amylovora* represents a benchmark in fire blight research. The sequence and its analysis will contribute to the formulation of new hypotheses and open the opportunity for new approaches to study this important pathogen. In the following years, we hope that the genome sequence will facilitate greater understanding of fire blight at both the molecular genetic and practical disease control levels.

MATERIALS AND METHODS

DNA Preparation and Sequencing

Strain Ea273 (ATCC 49946) of *E. amylovora* was isolated from an infected apple tree growing in a western New York orchard in the 1970s. Genomic DNA was extracted from cells grown from a lyophilized sample of Ea273. Sequencing was performed at The Sanger Institute. The initial readings were obtained by sequencing genomic shotgun libraries cloned in pMAQ1 using dye terminator chemistry in ABI3730 automated sequencers.

Sequence Analysis, Annotation and Phylogenetic Analysis

The sequence was assembled, finished and annotated using methods similar to those described previously (Parkhill et al., 2003). The finished sequence and the shotgun readings are publicly available at http://www.sanger.ac.uk/Projects/E_amylovora/ .

The sequence was analyzed with the following publicly available programs: Glimmer 3.02 (http://cbcb.umd.edu/software/glimmer/) to predict protein coding regions; tRNAscan-SE 1.21 (http://lowelab.ucsc.edu/tRNAscan-SE/) for tRNA predictions; RBSfinder (http://www.tigr.org/software/genefinding.shtml) to find Ribosomal Binding Sites in bacterial and archaeal genomes, and Transterm (http://transterm.cbcb.umd.edu/) to predict rho terminators. Searches for orthologous sequences of predicted Open Reading Frames (ORF) used blastp, tblastn, tblastx and psi-blast options of the BLAST tool

available at http://www.ncbi.nlm.nih.gov/BLAST/ with the genome sequence as a query.

Artemis (Rutherford et al., 2000) was used as the visualization and annotation tool. ORFs were defined as pseudogenes if they were interrupted by mutations that would prevent translation, i.e. frameshifts, or if they had large portions of DNA inserted or deleted, compared with their functional homologs.

The finished sequence was compared with the genome sequences of selected bacteria using the MAUVE software (Darling et al., 2004). The sequences were aligned using the progressive MAUVE option and visualized with the same tool. The finished genome and preliminary annotations were uploaded to ASAP where we continue to refine predictions and comparative analyses.

RESULTS AND DISCUSSION

Genome Structure and General Features

The genome of Ea273 consists of a single circular chromosome of 3,805,874 bp with a G+C content of approximately 53.5%. In addition, it has two extra-chromosomal plasmids, one of 71,487 bp and one of 28,243 bp. A comparison of the general features of the genome of Ea273 is presented in Table 1. Interestingly, the chromosome of Ea273 is the smallest of the genomes compared. The predicted coding sequences (CDS) covers approximately 85% of the genome and their average length is 939 bp; both values are similar to the mean of other closely related enterobacteria and the plant pathogen *Pseudomonas syringae* pv. tomato DC3000 (Pst) (Buell et al., 2003) (Table 1). The only exception is *Sodalis glossinidius* (SLG) which has a very low percentage of CDS, a large number of pseudogenes and shorter average length, characteristic of obligate symbionts. Similarly, the estimated number of tRNAs in the Ea273 genome is in the same range of those of the compared genomes. The number of predicted CDS is lower for Ea273 than for the other compared genomes except for SGL, but in contrast with the latter, the low number in Ea273 is due mainly to the shorter length of the chromosome and not because of a lower percentage of coding sequences. Thus, Ea273 has a more compact genome. The number of estimated pseudogenes for Ea273 is two and three times greater than in *Pectobacterium carotovorum* (ECA) (Bell et al., 2004) and *S. enterica* sv. *typhimurium* LT2 (SENT) (McClelland et al., 2001), respectively. These species are considered examples of host generalists, whereas *E. amylovora* is rather specific. However, when compared with highly specialized endoparasites, the number of pseudogenes of Ea273 is much smaller, i.e. only one-ninth that of *Y. pestis* CO92 (YPE) (Parkhill et al., 2001). This suggests that Ea273 is somewhat specialized for its niche while remaining capable of a free-living life style.

A total of 3376 CDSs were identified in the chromosome of Ea273; 292 (9%) do not have any match in the current NCBI databases, while 324 (10%) are conserved hypothetical proteins with no known function and only 46 (1%) seem to be mobile elements such as integrases, transposases or phage related CDSs. This number is low compared with other plant pathogenic bacteria, which also supports the notion that Ea273 has a stable genome. The remaining 80% of the CDSs were roughly classified in broad categories. Most genes (787 CDS or 29%), are involved in transport or thought to be associated with membranes, 18% are associated with cellular processes or energy production and 17% are associated with nutrition/metabolic functions. Type III and Type II secretion systems and their associated substrates, flagella and fimbriae biosynthesis genes were roughly grouped in an important category comprising 17% of the total genes with homology to other genes in the databases searched.

Plasmids

Two plasmids were identified in the genome of Ea273. The first plasmid is a 28.24 kbp and has been reported previously (McGhee and Jones, 2000) as pEa29. This sequence for Ea273 has 100% identity with the one previously reported. Our annotation predicts seven pseudogenes and four additional CDSs for this plasmid.

The second plasmid, 71.49 kbp, identified as pEa72 to follow previous nomenclature, has some similarity at the DNA level in its first 5500 nt with pEU30, reported previously (Foster et al., 2004) for strains of *E. amylovora* collected from Pacific coast states of the USA; the region from 5500 nt to ca. 30000 nt seems to be related to genes of enterobacterial origin. pEa72 contains 87 predicted CDSs, with only two predicted mobile/phage-related CDSs and one pseudogene. Forty-one % of the remaining 84 CDSs code for *tra*-like and *pil*-like conjugative transfer genes, while 33% have homology to hypothetical proteins of no known function or no matches in the databases. Thirteen % of the CDSs appear to be involved in replication and stability.

T3SS Systems

The known *hrp* pathogenicity island (Oh et al., 2005) was located in the genome of *E. amylovora.* In addition two novel T3SS islands were identified. They were named PAI2 and PAI3 to differentiate them from the known pathogenicity island (PAI1). Both regions (PAI2 and PAI3) encode complete sets of T3SS apparatus proteins. When individual translated proteins in each region were compared with the REFSEQ database (BLASTP), the structures and organizations of both islands were very similar (Fig. 1), and they were most similar to the SSR-1 island of SGL, which has two different functional T3SS islands (Dale et al., 2005). SSR-1 is similar to the *ysa* chromosomal island of *Y. enterocolitica* (Foultier et al., 2002), which is required for invasion of host cells by both organisms (Dale et al., 2005).

Due to the similarity in structure and organization of PAI2 and PAI3, and the completeness of PAI3 with respect to PAI2, we speculate that PAI2 might be paralogous of PAI3. While PAI1 clearly is required for pathogenicity of Ea273 in its plant host (Steinberger, 1988), the function of PAI2 and PAI3 is not apparent. These islands encode genes that seem more related to animal pathogens or endosymbionts; however, no animal vectors or infected hosts of *E. amylovora* have been reported. Therefore, expression and functional studies must be undertaken to elucidate the role of the novel T3SSs of PAI2 and PAI3.

Comparative Genomics

We compared the genome of Ea273 with available genome sequences from six closely related species [SENT, *E. coli* O157:H7 (EHEC), *E. coli* K12, *P. carotovorum*, *E. chrysanthemi* and SGL] to the genome of Ea273 using the progressive Mauve option of the Mauve comparison software (Darling et al., 2004). In concordance with the phylogenetic relationship (Hauben et al., 1998) (Fig. 2), the three genomes with closer relationships are EHEC, SENT and the non-pathogenic *E. coli* K12. The progressive MAUVE comparison was repeated for only Ea273, SENT and ECOK12 to determine more clearly the backbone sequence. The detailed analysis of the three sequences (Fig. 2) suggests that the loss of putative coding regions occurred throughout all the genomes, as we observed a high number of locally collinear blocks (LCB) shorter in Ea273 than in either ECOK12 or SENT. Some large rearrangements were evident, especially around the origin of replication. The analysis revealed that a large proportion of missing genes are related to environmental responses and to definite metabolic pathways, such as anaerobic respiration or sugar metabolism which coincide with previously reported deficiencies (Paulin, 2000).

Contributions of the Genome to the Fire Blight Community

The complete sequence of the genome of *E. amylovora* is publicly available at the Sanger web page (http://www.sanger.ac.uk/Projects/E_amylovora/). The reports of the genome sequence with the annotation, including a manual annotation for virulence factors; and a bioinformatics analysis that produced an inventory of candidate genes upregulated by HrpL, a sigma factor known to upregulate T3SS related genes during plant pathogenic infection in *E. amylovora* and other plant pathogens are in preparation for publication.

The complete sequence and annotations are currently available at the ASAP web page (https://asap.ahabs.wisc.edu/asap/logon.php) upon registration. Interested scientists can register by contacting Nicole T. Perna or Ana Maria Bocsanczy. Registered members of the scientific community can retrieve information concerning the genome and make comparisons, and/or they contribute to the ongoing annotation of the genome.

A web page on fire blight, which will be a portal to the genome project, K12 outreach programs, fire blight information, and other links of interest to the fire blight community, will be publicly available.

The availability of the *E. amylovora* genome opens a new era of post-genomics analysis, which offers an important source of information for functional and evolutive studies. Comparisons with other sequenced plant-pathogenic, which might have similar or very different mechanisms of virulence and broader host ranges, are now possible. The comparisons will also provide new sources of experimental information that might help elucidate the function of the homologous genes found in *E. amylovora*. Additionally, the availability of the genome opens new questions for specific functional studies that might impact present research, such as functional studies of translocation.

ACKNOWLEDGEMENTS

This work was supported by NSF/USDA CSREES Microbial Genome Sequencing Grant number 2004-35600-14258.

Literature Cited

Bell, K.S., Sebaihia, M., Pritchard, L., Holden, M.T.G., Hyman, L.J., Holeva, M.C., et al. 2004. Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* subsp. *atroseptica* and characterization of virulence factors. Proc. Natl. Acad. Sci. USA 101:11105–11110.

Buell, C.R., Joardar, V., Lindeberg, M., Selengut, J., Paulsen, I.T., Gwinn, M.L., et al. 2003. The complete genome sequence of the Arabidopsis and tomato pathogen *Pseudomonas syringae* pv. tomato DC3000. Proc. Natl. Acad. Sci. USA 100:10181–10186.

Dale, C., Jones, T. and Pontes, M. 2005. Degenerative evolution and functional diversification of type-III secretion systems in the insect endosymbiont *Sodalis glossinidius*. Mol. Biol. Evol. 22:758–766.

Darling, A.C.E., Mau, B., Blattner, F.R. and Perna, N.T. 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Research 14:1394–1403.

Foster, G.C., McGhee, G.C., Jones, A.L. and Sundin, G.W. 2004. Nucleotide sequences, genetic organization, and distribution of pEU30 and pEL60 from *Erwinia amylovora*. Appl. Environ. Microbiol. 70:7539–7544.

Foultier, B., Troisfontaines, P., Muller, S., Opperdoes, F.R. and Cornelis, G.R. 2002. Characterization of the ysa pathogenicity locus in the chromosome of *Yersinia enterocolitica* and phylogeny analysis of type III secretion systems. J. of Molecular Evolution 55:37–51.

Glasner, J.D., Liss, P., Plunkett, G., Darling, A., Prasad, T., Rusch, M., et al. 2003. ASAP, a systematic annotation package for community analysis of genomes. Nucleic Acids Res. 31:147–151.

Hauben, L., Moore, E.R.B., Vauterin, L., Steenackers, M., Mergaert, J., Verdonck, L., et al. 1998. Phylogenetic position of phytopathogens within the Enterobacteriaceae. Syst. Appl. Microbiol. 21:384–397.

McClelland, M., Sanderson, K.E., Spieth, J., Clifton, S.W., Latreille, P., Courtney, L., et al. 2001. Complete genome sequence of *Salmonella enterica* serovar *typhimurium* LT2. Nature 413:852–856.

McGhee, G.C. and Jones, A.L. 2000. Complete nucleotide sequence of ubiquitous plasmid pEA29 from *Erwinia amylovora* strain Ea88: Gene organization and intraspecies variation. Appl. Environ. Microbiol. 66:4897–4907.

Momol, M.T. and Aldwinckle, H.S. 2000. Genetic Diversity and Host Range of *Erwinia amylovora*. p.55–72. In: J.L. Vanneste (ed.), Fire Blight: The Disease and Its Causative Agent, *Erwinia amylovora*. Wallingford, Oxon, UK; New York, NY, USA. CABI Pub.

Oh, C.S., Kim, J.F. and Beer, S.V. 2005. The Hrp pathogenicity island of *Erwinia amylovora* and identification of three novel genes required for systemic infection. Mol. Plant Pathol. 6:125–138.

Parkhill, J., Sebaihia, M., Preston, A., Murphy, L.D., Thomson, N., Harris, D.E., et al. 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. Nat. Genet. 35:32–40.

Parkhill, J., Wren, B.W., Thomson, N.R., Titball, R.W., Holden, M.T.G., Prentice, M.B., et al. 2001. Genome sequence of *Yersinia pestis*, the causative agent of plague. Nature 413:523–527.

Paulin, J.-P. 2000. *Erwinia amylovora*: general characteristics, biochemistry and serology. p.87–116. In: J.L. Vanneste (ed.), Fire Blight: The Disease and Its Causative Agent, *Erwinia amylovora*. Wallingford, Oxon, UK; New York, NY, USA. CABI Pub.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., et al. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16:944–945.

Steinberger, E.M. and Beer, S.V. 1988. Creation and complementation of pathogenicity mutants of *Erwinia amylovora*. Molecular Plant-Microbe Interactions 1:135–144.

Tables

Table 1. General characteristics of the genome of Ea273 as compared with selected bacterial genomes.

		Erwinia amylovora	*E. coli* K12	*E. coli* O157:H7	*Salmonella enterica* subsp. *typhimurium*	*Sodalis glossinidius*	*Pectobacterium carotovorum*	*Pseudomonas syringae* pv. tomato
Strain		Ea273	MG1655	EDL933	LT2		SCRI1043	DC3000
Size	Kbp	3805.9	4639.2	5528.4	4857.4	4171.1	5064	6397.1
G+C content	%	53.6	50.8	50.5	53.0	54.0	51.0	58.4
Coding regions	%	85.1	87.8	86.2	86.0	50.9	85.9	86.8
Average CDS length	bp	939	952	903	903	873	973	988
Plasmids		2	1	2	1	3	0	2
Predicted tRNAs		78	86	98	85	69	76	63
Predicted CDSs		3376	4288	5340	4450	2432	4439	5615
Pseudogenes		110	27	6	39	972	52	

Figures

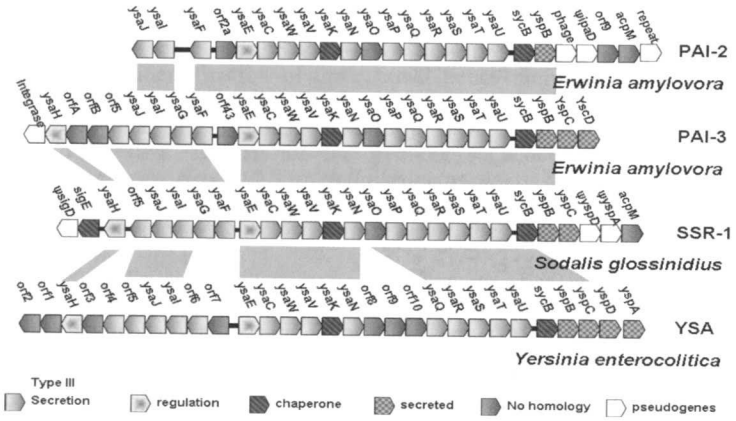

Fig. 1. Comparative organization and similarity of the novel T3SS islands. The genetic organization of PAI-2 and PAI-3 are very similar to SSR-1 of *S. glossinidius* and *Y. enterocolitica*. The figure was adapted from Dale et al. (2004). Each gene is indicated with an arrow, the gray shades and pattern codes are indicated in the lower row. The gray blocks indicate the regions of synteny.

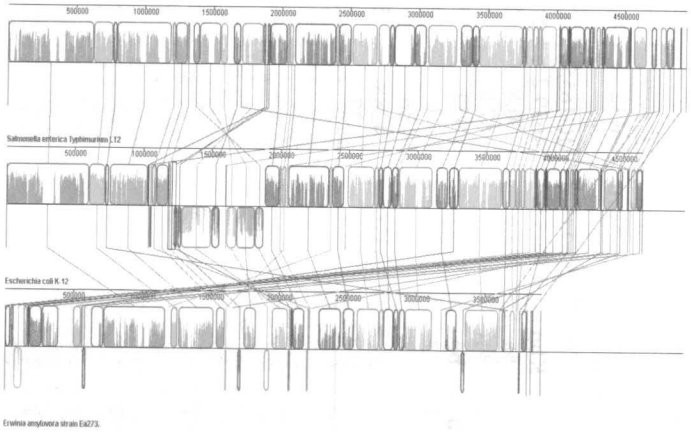

Fig. 2. Comparative genomics of three genomes. From upper to lower row: *Salmonella enterica* subsp. *thyphimurium*, *Escherichia coli* K12 and *Erwinia amylovora* Ea273. The alignment was produced and visualized with the progressive Mauve option of the Mauve comparison software (Darling et al., 2004). The blocks represent locally collinear blocks in the three genomes (LCB). The lines inside each block represent the similarity of the regions inside the blocks.